ISSN 2395-1621

Consumer Product Recommendation System based on Text and Context Mining

[#1]Chetankumar Madane, [#2]Shreya Gadikar, [#3]Anand Deshmukh,
[#4]Himanshu Kumar, [#5]Prof. S.A.Joshi

[3]ananddeshmukh177@gmail.com

[#12345]Department of Computer Engineering,

Sinhagd College of Engineering, Vadgaon(Bk.),
Pune, Maharashtra ,India

ABSTRACT

Business Analysis (BA) is a concept of applying a set of technologies to convert data into meaningful information. Large amounts of data originating in different formats and from different sources can be consolidated and converted to key business knowledge. Modeling and analysis are two critical steps in any process redesign effort. Data mining and Business analysis work together to process data and analyze it in a way that eases the workload for the users and aids with the understanding of the materials. Business analysis focuses more on data integration and simulation of data. The organization current issue, the new breed of technique is required that has mining the data reviews and classify to its ratings. The design of such a hierarchical simulation tool called business analysis tool. Dependence of users on web applications is increasing very rapidly in recent time, hence Star based ratings are used in these product sites. But existing system is unable to make proper use of the customer's given comments. Prediction and Classification of web application contains brand wise sorting of products but feature wise ratings and sorting is not available. This system will give textual classification based on customer's comments of particular product. This will enable the feature wise ratings and accurate sorting of the product which aims towards better user experience and increased product sales.

ARTICLE INFO

Article History

Received: 3rd June 2018

Received in revised form :

3rd June 2018

Accepted: 5th June 2018

Published online :

8th June 2018

Keywords: Business Analysis , Modelling ,textual classification.

I. INTRODUCTION

Dependence of web applications on day-to day users is increasing with time. For example ,most of the customers who wish to buy a product look for online shopping since it is easier than there manual shopping. Hence to even further improve the user-experience by providing ratings converted from text based reviews given by users who bought the product. Choosing a particular product for a specific user has always been a difficult and time consuming task .Customers always have to search through the user reviews forth perfect description of the product they need to purchase. This has been solved upto an extent with the current star based ratings. But that doesn't work while choosing star base ratings. But that doesn't work while choosing a particular feature in a production the existing system.

II. PROPOSED SYSTEM

This project considers textual based comments as well as numeric based ratings for recommendation. This project uses pre-processing, Naive Bayes and decision-making system to classify the textual comments. This project will provide category wise, feature wise, brand wise sorting along with textual-comment wise sorting and hit wise sorting . Instead of admin reading all the comments and making decision ,this system will give decision making on the user comments of the particular product.

A. Fundamentals

Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is an essential process where intelligent methods are applied to extract data patterns. It is an interdisciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and

transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations post processing of discovered structures, visualization and online updating. Data mining is the analysis step of the "Knowledge discovery in database" process, or KDD.

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but do belong to the overall KDD process as additional steps.

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

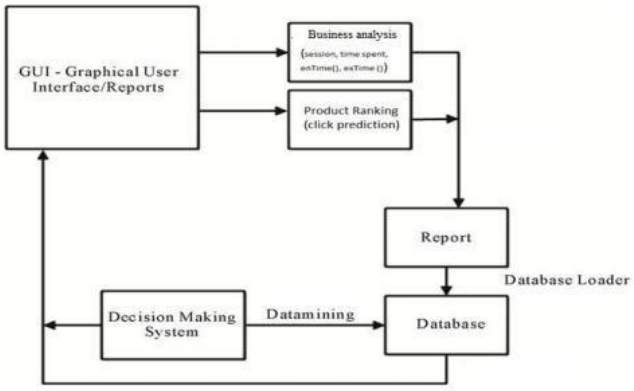

Figure 1. Architecture diagram

B. System Overview

An architectural overview of the project is illustrated above. The GUI will contain all the retrieval methods for the user comments. It will also have the ability to view the ranking of the products, and also the recommendations based on the user preferences. After the retrieval of the comment, a report will be generated. This report will then be loaded into the database using database loader. Then the decision making system will come into picture. The decision making system will then make use of the Naive Bayes algorithm to specifically identify and rate the good and the bad ratings accordingly. This algorithm may refer to the database for the dictionary for the meaning of the words prescribed into the comments retrieved from the GUI. After the ratings are assigned to the particular comment, the comment, its ratings, and also the updated rating of the product or the feature mentioned will be displayed through the GUI.

C. Survey

Machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s, and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihoodtraining can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features. It is based on the Bayes theorem which describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if cancer is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have cancer, compared to the assessment of the probability of cancer made without knowledge of the person's age. It is used in this project and can be formulated as.

$$p(C_k \mid x) = (p(C_k)\, p(x \mid C_k)) / p(x)$$

III. METHODOLOGY

The web technologies used will be JSP, and JavaScript. Java Server Pages (JSP) is a technology that helps software developers create dynamically generated web pages based on HTML, XML, or other document types. Released in 1999 by Sun Microsystems, JSP is similar to PHP and ASP, but it uses the Java programming language. To deploy and run JavaServer Pages, a compatible web server with a servlet container, such as Apache Tomcat or Jetty, is required. Hence the server used will be Apache Tomcat server. Apache Tomcat, often referred to as Tomcat Server, is an open-source Java Servlet Container developed by the Apache Software Foundation (ASF). Tomcat implements several

Java EE specifications including Java Servlet, JavaServer Pages (JSP), Java EL, and WebSocket, and provides a "pure Java" HTTP web server environment in which Java code can run. It has also added user- as well as system-based web applications enhancement to add support for deployment across the variety of environments. It also tries to manage sessions as well as applications across the network.

Java is used because it is easy to implement onto the web applications, and it will also be easier to run and maintain since most of the web applications today are based on Java Server Pages. The Database used is MySQL. MySQL is an open-source relational database management system(RDBMS). Its name is a combination of "My", the name of co-founder Michael Widenius's daughter, and "SQL", the abbreviation for Structured Query Language. The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL was owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Oracle Corporation. For proprietary use, several paid editions are available, and offer additional functionality.

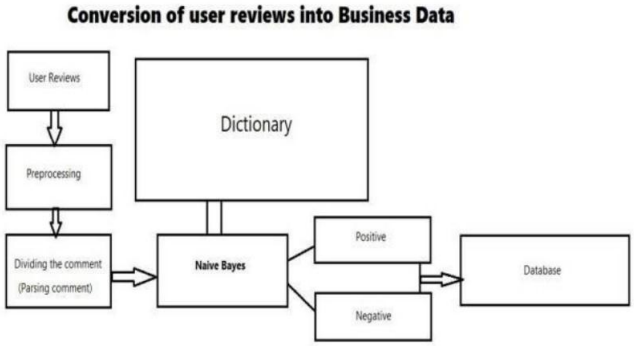

Figure 2. Rating and classification

MySQL is open source and also supports SSL support, Query Catching, and also Embedded database library which greatly improves the user experience that this project is able to provide.

IV. RESULT

The following results are obtained:

1.We created a table which is often used to describe the performance of classification model called as confusion matrix.

2.Calculated the quality condition or fact of being accurate also called as precision.

3.Also Calculated sensitivity of system called as recall.

Confusion Matrix

	Predicted Negative	**Predicted Positive**
Negative Cases	**True Negative**	**True Positive**
Positive Cases	**False Negative**	**False Positive**

TP=30 ,TN=20, FP=8, FN=5

	Predicted Negative	**Predicted Positive**
Negative Cases	**20**	**30**
Positive Cases	**5**	**8**

Precision : TP/(TP+FP)
Precision : 30/(30+8)
Precision : 0.7894
Recall : TP/(TP+FN)
Recall : 30/(30+5)
Recall : 0.8571

V. CONCLUSION

Online shopping and the general use of web applications is an important part in the day to day life of the regular customer since anyone can shop and pay while not being required to leave their respected homes. The user experience comes to a great functionality while consider these types of web applications. Hence to further improve the user-experience of the average consumer, this project aims to provide a better recommendation system for the interested consumers while also providing the quick numerical ratings of the already existing and new textual reviews in the form of comments.

REFERENCES

[1] Shahab Saquib Sohail, Jamshed Siddiqui, Rashid Ali, "Feature extraction and analysis of online reviews for the recommendation of books using opinion mining technique", Science Direct 2017.

[2] Shahab Saquib Sohail, Jamshed Siddiqui, Rashid Ali," User feedback scoring and Evolution of a product Recommendation system",IEEE 2014.

[3] Aimin Luo, Jiong Fu, Junxian Liu," An Impact Analysis Method of Business Processes Evolution in Enterprise Architecture", IEEE 2016

[4] Pierre-Aymeric ,Masse and Nassim Laga," A Contextual Data Selection Tool for an Enhanced Business Process Analysis"

[5] Jingjing Cao, Wenfeng Li, " Sentimental feature based Collaborative Filtering Recommendation", IEEE 2016

[6]M. Lankhorst,et al, "Enterprise architecture at work: Modelling,Communication and analysis", Third Edition, Springer-Verlag BerlinHeidelberg 2013.

[7]Aciar, S., et al,2006.Recommender System based on Consumer product reviews. In: proceedings of the 2006 IEEE/WIC/ACM International conference on web intelligence. IEEE computer society.

[8]H. U. ,M. , Liu, B., 2004 Mining and Summarizing customer reviews. In: KDD'04.

[9] Sohail, S. S. Siddiqui, J. , Ali, R.2014b, User feedback scoring and evaluation of a product recommendation system. In: proceedings of IC3, PP.525-530.

[10] Teng, Z., Ren, F., & Kuriowa, S. (2007). Emotion recognition from text based on the rough set theory and the support vector machines. In 2007 international conference of the natural language processing and www.ierjournal.org knowledge engineering (pp. 36– 41). Beijing, China: IEEE Computer Society.

